Dr Velibor ILIĆ, naučni saradnik,
Istraživačko-razvojni institut za veštačku inteligenciju Srbije
Msr Lenka BAJČETIĆ, doktorand
Dr Snežana PETROVIĆ, naučni savetnik,
Institut za srpski jezik SANU
Dr Ana ŠPANOVIĆ, naučni saradnik,
Institut za srpski jezik SANU

SCyDia – OCR. SOFTVER ZA SRPSKU ĆIRILICU SA DIJAKRITICIMA

Apstrakt: U tekućem procesu digitalizacije srpskih dijalekatskih rečnika, najveća prepreka je nedostatak mašinski čitljivih verzija papirnih izdanja. OCR obrada nije nova tehnologija, jer postoji mnogo softverskih rešenja otvorenog koda i komercijalnih rešenja koja mogu pouzdano da konvertuju skenirane slike papirnih dokumenata u digitalne dokumente. Dostupna softverska rešenja obično su dovoljno efikasna za obradu skeniranih ugovora, faktura, finansijskih izveštaja, novina i knjiga. Međutim, u slučajevima kada je potrebno obraditi dokumente koji sadrže akcentovani tekst i precizno izdvojiti svaki znak sa dijakritičkim obeležjima, ovakva softverska rešenja nisu dovoljno efikasna. U ovom radu predstavljamo OCR softver pod nazivom SCyDia, razvijen da prevaziđe ovakve probleme. U radu su prikazani organizaciona struktura OCR softvera SCyDia i prvi rezultati obrade. SCyDia je softversko rešenje zasnovano na vebu koje se oslanja na softver otvorenog koda Tesseract u pozadini. SCyDia takođe sadrži modul za poluautomatsku korekciju teksta. Do sada smo obradili preko 15.000 stranica, posebno trinaest dijalekatskih rečnika i pet dijalekatskih monografija. U ovom trenutku analizirali smo tačnost SCyDia-a obradom trinaest dijalekatskih rečnika. Rezultate su ručno analizirali stručnjaci koji su pregledali više nasumično odabranih stranica iz svakog rečnika. Preliminarni rezultati pokazuju tačnost u prepoznatom tekstu koja se kreće u opsegu od 97,19% do 99,87%.

Ključne reči: *OCR, ćirilična slova, retrodigitalizacija, konvolutivne neuronske mreže, CNN*

UVOD

Na Institutu za srpski jezik SANU trenutno je u toku nekoliko leksikografskih projekata – deskriptivnih, etimoloških, istorijskih, dijalekatskih i dr., koji se sastavljaju na tradicionalan način. Leksička građa na kojoj se zasnivaju obuhvata brojne rečnike i naučne monografije koje je potrebno konsultovati u papirnom izdanju. Ogromna većina ovih rečnika i monografija (na desetine hiljada stranica), posvećenih sastavljanju i analizi dijalekatske leksike i opisu dijalekatske osobenosti, pisana je ćirilicom, sa akcentima, dijakritičkim i drugim nestandardnim znakovima. Treba imati u vidu da je srpski jezik u poziciji malog resursa u oblasti digitalne infrastrukture i digitalizovanih jezičkih resursa (npr. u Institutu nijedan rečnik nije korpusno zasnovan niti korpusno vođen i sl.). Uprkos činjenici da su učinjeni ozbiljni prvi koraci u primeni novih tehnologija ka našem leksikografskom nasleđu i u procesu izrade rečnika, bili smo svesni da ova zastarela

*Рад је презентован као уводно предавање на Саветовању „Куршумлијска Бања 2023“.

metodologija može dovesti u pitanje relevantnost rezultata istraživanja i umanjiti naučni nivo publikacija. Zbog toga smo odlučili da zauzmemo širi pristup kako bismo unapredili naš rad – da retrodigitalizujemo ovaj ogroman broj naučnih rečnika i monografskih studija od fundamentalnog značaja za leksikografski rad. To će nam omogućiti da napravimo multifunkcionalnu leksikografsku bazu podataka i različite korpuse, kao i koristimo dijalekatski materijal za izradu raznih rečnika i naučnih radova, što će istovremeno biti promocija i dijalekata i narodnih jezika. Najveća prepreka u pokušaju retrodigitalizacije srpskih dijalekatskih rečnika bio je nedostatak mašinski čitljivih verzija papirnih izdanja, što je podrazumevalo da moramo da uradimo osnovni korak pre nego što se upustimo u proces izrade rečnika u digitalnom okruženju – OCR-ovanje stranica sa najvećom mogućom tačnošću.

Optičko prepoznavanje znakova (OCR) je proces koji omogućava ekstrakciju podataka iz skeniranog dokumenta ili slike. U ovom procesu, odštampani ili rukom pisani tekst na skeniranom dokumentu se pretvara u mašinski čitljiv format. OCR obrada nije nova tehnologija, pošto postoji veliki broj open-source i komercijalnih softverskih rešenja koja mogu pouzdano da konvertuju skenirane slike papirnih dokumenata u digitalne dokumente. Međutim, dostupna softverska rešenja su obično dovoljno efikasna za obradu skeniranih ugovora, faktura, finansijskih izveštaja, novina i knjiga, ali ne i u slučajevima kada je potrebno obraditi dokumente koji sadrže akcentovani tekst i precizno izdvojiti svaki znak sa dijakritičkim obeležjima, na primer dijalekatske rečnike pisane ćiriličnim slovima.

Zašto OCR?

Iako je dvostruka provera najtačniji način za transkripciju, ovakav proces je veoma dugotrajan, a u slučaju dijalekatskih i istorijskih rečnika, sa tekstom koji je previše složen za osobe koji nisu eksperti, proces je veoma skup, jer zahteva dodatne ispravke, obično više od jedne. Stoga smo, da bismo prevazišli ovaj problem, odlučili da investiramo u razvoj OCR softvera pod nazivom SCyDia – srpska ćirilica sa dijakritikom. Do sada smo pokrenuli softver SCyDia na četrnaest rečnika i monografija sa više od 15.000 stranica zajedno, ali nameravamo da ga koristimo na stotinama hiljada stranica više.

Pošto tačnost OCR varira od 97,19% do 99,87%, postoje neki rečnici koji se mogu prilično brzo ručno proveriti. S druge strane, najgori procenat greške od 2,81% koji je primećen na jednom rečniku znači da na stranici od 3.000 karaktera ima 84,3% grešaka, što može biti dugotrajno i preskupo za ispravljanje. U ovakvim slučajevima odlučili smo se za postepeni pristup tako što smo u prvoj fazi ispravili samo leme iz naslovne reči, jer bismo na ovaj način mogli da omogućimo pretraživanje baze podataka.

Pregled literature

Klyshinsky/Karpi/Bondarenko (2020) predstavlja rezultate poređenja softvera neuronske mreže koji se koriste za vraćanje dijakritičkih znakova na jezicima kao što su hrvatski, slovački, rumunski, francuski, nemački, letonski i turski, a tačnost prepoznavanja obično se kreće od 95-99% u zavisnosti od pisma (neka slova imaju nižu tačnost). Hussain/Niazi/Anjum/Irfan (2014) predstavlja rezultate korišćenja *Tesseract* motora za OCR obradu stranica koje je napisao Urdu Nastalique (veoma složen i kurzivni

stil pisanja arapskog pisma). Bez ikakvih modifikacija *Tesseract* postiže tačnost od 66%, a uz dodatne modifikacije tačnost je povećana na 97%.

Cristea/Pădurariu/Rebeja/Onofrei (2020) predstavlja rezultate rešenja zasnovanog na nekoliko tipova neuronskih mreža (kao što su *The Region Proposal Netvork* (RPN), *ResNet, Faster R-CNN*) za OCR obradu starih rumunskih dokumenata napisanih na ćirilici.

Rijhwani/Anastasopoulos/Neubig (2020) opisuje metode naknadne korekcije gde je cilj smanjenje broja grešaka koje se javljaju tokom OCR obrade, a koje se najčešće javljaju usled lošeg kvaliteta skeniranja, fizičkog pogoršanja papirne knjige ili različitih stilova fonta.

Krstev/Stanković/Vitas (2018) u svom istraživanju predstavljaju proces obnavljanja dijakritike u srpskim tekstovima pisanim degradiranim latiničnim pismom, a predstavljeno rešenje se oslanja na sveobuhvatne leksičke resurse za srpski jezik: morfološke elektronske rečnike, Korpus savremenog srpskog i lokalne gramatike.

O'Brien/Haddej (2012) u svom istraživanju predstavljaju projekat u kojem je proširena funkcionalnost softvera *OCRopus* da podrži prepoznavanje matematičkih simbola i jedinstvenih jezičkih alfabeta (npr. mađarska slova), dok proširena verzija podržava UTF-8 kodiranje karaktera. Tačnost originalne verzije obučene samo sa engleskim znakovima bila je 86%, a u proširenoj verziji tačnost je povećana na 93,5%.

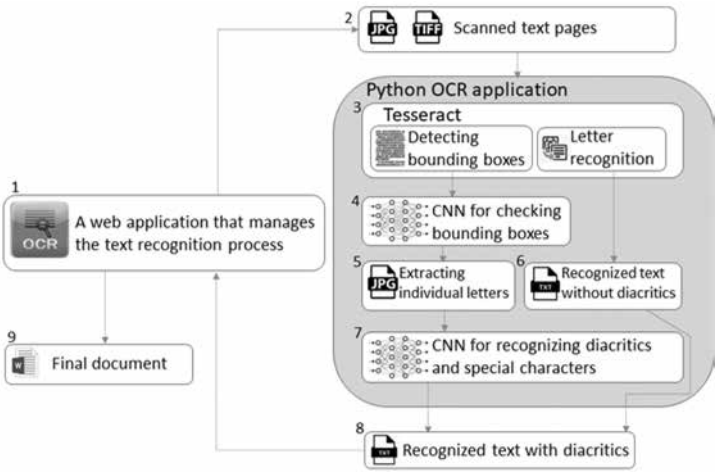
Pregled softvera SCyDia

U ovom radu ćemo predstaviti OCR softver SCyDia, softversko rešenje zasnovano na vebu koje se oslanja na softver otvorenog koda Tesseract u pozadini, razvijen da prevaziđe problem nedovoljno efikasnog OCR softvera za obradu dokumenata koji sadrže akcente teksta i da precizno izdvoji svaki znak sa dijakritičkim obeležjima. Prikazaćemo organizacionu strukturu softvera i prve rezultate.

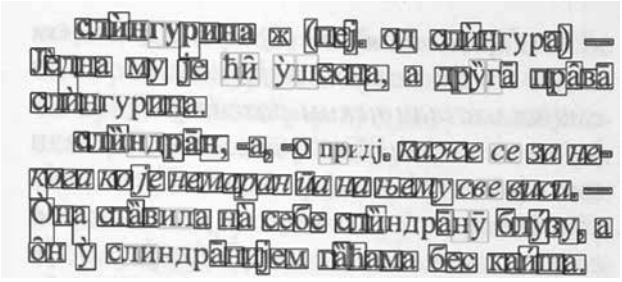
Rad je organizovan na sledeći način: Poglavlje 2 sadrži detalje implementacije, detalje o korišćenim konvolutivnim neuronskim mrežama (CNN) i skupovima podataka za treniranje mreže, kao i opis modula za poluautomatsku korekciju teksta. Nakon toga, u Poglavlju 3 su prikazani rezultati. Dalji planovi su predstavljeni u Poglavlju 4, dok poslednji deo sadrži zaključke.

IMPLEMENTACIJA SCYDIA SOFTVERA

SCyDia OCR softver je razvijen kao veb aplikacija (pregled algoritma predstavljen je na Slici 1) koja korisniku omogućava da vidi listu skeniranih stranica, odabere stranice za OCR obradu ili za ispravku teksta (lektorisanje). Veb aplikacija (1) omogućava korisniku da izabere koje će skenirane stranice biti obrađene. Izabrane slike skeniranih tekstualnih stranica (2) se prosleđuju u Python skriptu za obradu. OCR obrada u početnom koraku koristi Tesseract (3) koji generiše tekstualni fajl (6) sa prepoznatim tekstom bez dijakritičkih znakova. Tesseract može da generiše koordinate graničnih okvira oko pojedinačnih slova. Koordinate graničnih kutija obično su konkretno određene, mada se povremeno može desiti da se umesto jednog slova unutar graničnog okvira nađu i dva, tri ili čak više slova. Granični okvir ponekad može da sadržati polovine dva susedna slova.



Slika 1: *Pregled softvera SCyDia*



Slika 2: *Detektovani granični okviri oko slova*

Konvolutivna neuronska mreža (4) može da proveri da li granični okvir sadrži samo jedno slovo kao što je očekivano, a ako ima više slova vraća informaciju o tome koliko se njih nalazi unutar graničnog okvira. Detektovani granični okviri sa više od jednog slova mogu se podeliti na odgovarajući broj manjih graničnih okvira, od kojih svaki sadrži jedno slovo. Na Slici 2 plavom bojom su prikazani pravilno određeni granični okviri sa jednim slovom, zelenom bojom okviri koji su prvobitno sadržali dva slova i koji su podeljeni na dva dela, a okviri koji su podeljeni na tri slova prikazani su žutom bojom.

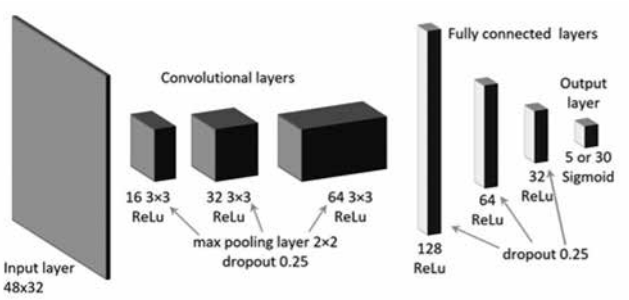
Na Slici 1 *Python* skripta (5) koristi koordinate graničnih okvira za izdvajanje slika pojedinačnih slova. Konvolutivna mreža (7) obrađuje slike pojedinačnih slova i pokušava da otkrije da li sadrže dijakritičke znake. Takođe, ova mreža se može koristiti za prepoznavanje onih slova koje Tesseract ima poteškoća da pravilno prepozna, kao što su kurzivna. U poslednjem koraku, Python funkcija pokušava da upari svako pojedinačno slovo iz tekstualne datoteke sa informacijama koje pruža konvoluciona mreža prilikom obrade izdvojenih slika pojedinačnih slova. Rezultat ove funkcije predstavlja novi tekstualni fajl koji sadrži slova sa dijakritičkim znakovima.

Konfiguracija konvolutivne mreže i skupovi podataka SCyDia

OCR aplikacija koristi dve konvolucione neuronske mreže – CNN za proveru graničnih okvira i CNN za detekciju dijakritičkih znakova. Ove dve mreže imaju slične konfiguracije, a razlikuju se po broju izlaza.

CNN mreža koja se koristi za dijakritičku detekciju uzima matricu 48×32×1 kao ulaz i sadrži tri konvolutivna sloja. Prvi sloj sadrži 16, drugi 32, a treći 64 3×3 jezgra sa ReLu aktivacijom, na svaki od ovih slojeva postavlja se *max pooling* sloj dimenzija od 2×2,

a verovatnoća *dropout*-a je 0,25. Ove konvolucione slojeve prate potpuno tri povezana sloja: prvi sadrži 128 noda, drugi 64 noda, a treći 32 izlazna noda. Nakon svakog od ovih slojeva postavlja se sloj sa *dropout*-om od 0,25. I konačno, izlazni sloj sadrži 30 čvorova (Slika 3). Vrednosti dobijene na izlazu mreže imaju sledeće značenje: prva vrednost ukazuje da li slovo sadrži dijakritičke znakove, druga da li je slovo ispravno (ponekad granični okvir oko slova nije validan), sledećih 15 vrednosti detektuje tip dijakritičkih znakova, dok se preostale vrednosti koriste za otkrivanje slova koja Tesseract ne detektuje ispravno, na primer slova ьəђŋʒ i kurzivna slova kao što su: *ī ū ū̄ в*



Slika 3: *Struktura konvolutivnih mreža*

Dataset za treniranje CNN koji se koristi za detekciju dijakritičkih znakova generiše se prikupljanjem isečenih pojedinačnih slova sa skeniranih stranica.

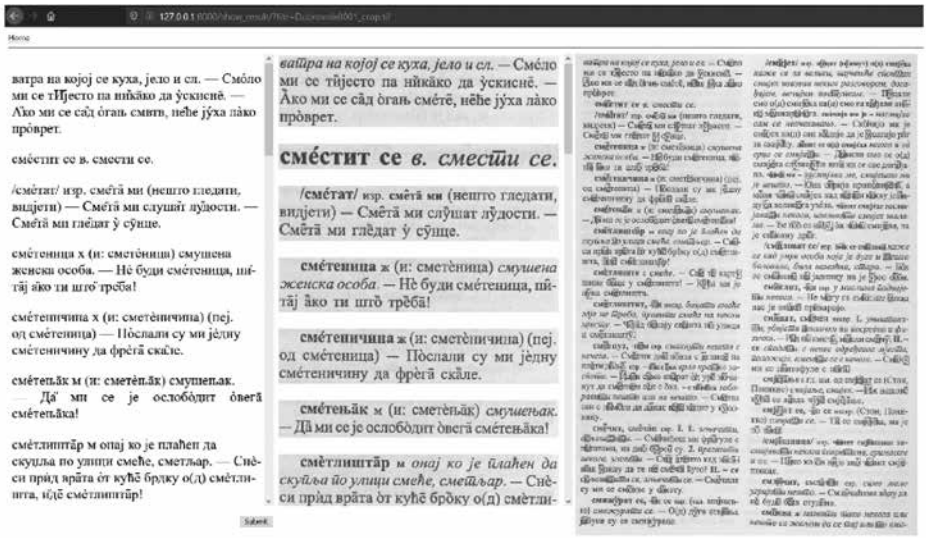
Ovaj skup podataka sadrži sledeće grupe ćiriličnih slova:

- standardni set ćiriličkih slova;
- slova koja imaju dijakritičke znakove iznad, na primer: à á ầ â ā ả ắ ã ấ ắ ả
- slova koja imaju dijakritičke znakove ispod, na primer: ạ a̤ ậ д̣
- slova koja imaju dijakritičke znakove iznad i ispod;
- ćirilična slova koja ne pripadaju standardnom skupu simbola koje Tesseract ne može da prepozna, na primer: ь ə ђ ŋ ʒ Tesseract pogrešno prepoznaje ova slova kao: Б о ђ о з
- slova u kojima se jedno slovo sastoji od dva znaka, na primer: д͡з

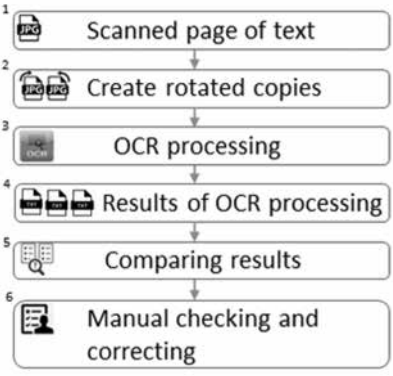
CNN mreža koja se koristi za proveru graničnih okvira ima sličnu konfiguraciju – izlazni sloj te mreže sadrži četiri noda (Slika 3). Vrednosti koje se dobijaju na izlazu mreže imaju sledeće značenje: prva ukazuje da li se u graničnom okviru nalazi samo jedno slovo, druga da se granični okvir nalazi oko dva slova, treća označava da je granični okvir oko tri slova, a četvrta se koristi za otkrivanje nevažećih slova (na primer unutar graničnog okvira postoje dve polovine uzastopnih slova). Skup podataka za CNN koji se koristi za proveru graničnih okvira takođe se generiše prikupljanjem isečenih slova sa skeniranih stranica. Ovaj dataset sadrži primere kako izgleda pravilno isečeno slovo, primere kada su dva ili tri slova zajedno izdvojena i primere slika sa pogrešno izdvojenim slovima, tj. kada se u graničnom okviru nalaze dve polovine slova. Adam optimizator se koristi prilikom treniranja obe mreže. Trajanje obuke je ograničeno na 50 epoha, sa dva dodatna parametra: *ReduceLROnPlateau* sa strpljenjem 10 i *EarlyStopping* sa strpljenjem 25. Parametar *ReduceLROnPlateau* smanjuje stopu učenja ako nije bilo poboljšanja tačnosti skupa podataka validacije za 10 epoha. *EarlyStopping* prekida obuku ako nema poboljšanja tačnosti skupa podataka validacije za 25 epoha.

Ручна и poluautomatska korekcija teksta (lektorisanje)

Osnovna namena aplikacije SCyDia je OCR obrada, a pored toga veb aplikacija obezbeđuje i modul za ispravljanje teksta (lектorisanje). Taj modul omogućava ručnu i poluautomatsku korekciju teksta. Prozor za ručnu ispravku teksta podeljen je na tri polja (Slika 4): prvo sadrži prepoznati tekst i to je polje za uređivanje, drugo sadrži isečene slike pasusa, u trećem je kompletna slika skenirane stranice na kojoj su označena slova koja sadrže dijakritičke znakove.



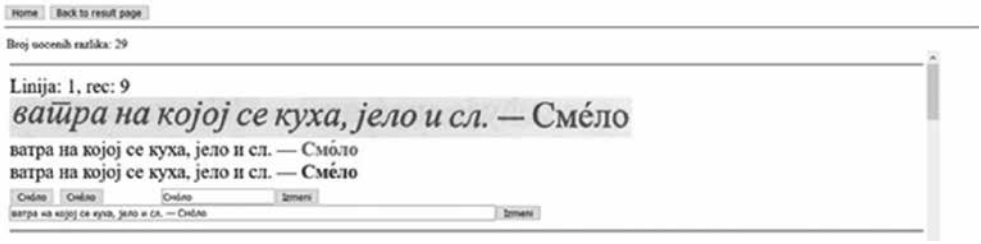
Slika 4: *Prozor za ručnu korekciju teksta*



Slika 5: *Algoritam za poluautomatsku korekciju teksta*

Da bi se postigla poluautomatska korekcija teksta (Slika 5), aplikacija SCyDia ponavlja OCR obradu (3) jedne stranice nekoliko puta kako bi kreirala dodatne kopije tekstualnih datoteka koje se mogu uporediti jedna sa drugom. Algoritam za poluautomatsko ispravljanje teksta počinje kreiranjem dodatne dve kopije (2) skenirane stranice (1) – prva slika se rotira ulevo za pola stepena, a druga kopija udesno za pola stepena, eventualno stepen. Ljudi neće primetiti razlike između originalne skenirane stranice i kopija te slike rotirane za pola stepena ako ih vizuelno uporede, ali za OCR softver tako mala razlika uzrokuje da se pogrešno prepoznata slova pojavljuju na različitim mestima u prepoznatom tekstu.

U sledećem koraku te tekstualne datoteke (4) se međusobno upoređuju (5) i za svaku otkrivenu razliku prikazuju se na prozoru za ručnu proveru i ispravljanje (6). U većini slučajeva korisnici mogu samo da kliknu na dugme sa tačnom verzijom reči (Slika 6).



Slika 6: *Korisnički interfejs sa rezultatima poluautomatske korekcije teksta*

Korisnički interfejs sa rezultatima poluautomatske korekcije teksta sadrži sledeće elemente: deo skenirane slike sa linijom teksta gde se primećuje razlika, red teksta gde se primećuje razlika u odnosu na originalno skeniranu stranicu (reč gde je razlika predstavljena crvenom bojom), red teksta gde se primećuje razlika od rotirane kopije (reč gde je razlika predstavljena crvenom bojom), dugme sa verzijom reči iz prve datoteke, dugme sa verzijom reči iz druge datoteke i okvir za tekst koji korisniku omogućava da ručno ispravi grešku ako nijedna od ove dve verzije nije tačna.

Upotreba SCyDia

Aplikacija SCyDia je do sada korišćena za obradu preko 15.000 stranica dijalektičkih rečnika srpskog jezika. OCR proces se sprovodi na računaru Intel I9 sa procesorom od 12-jezgara i sa NVidia GeForce RTX 2070 SUPER grafičkom karticom. Aplikacija SCyDia može da obrađuje osam stranica paralelno, a svaka stranica se analizira tri puta: prvo u svom originalnom obliku, zatim se slika ukosi za pola stepena ulevo a potom udesno. U proseku je za obradu svake stranice potrebno oko pola sata. Nakon što je obrađena prva serija od 14 rečnika, analizirani su postignuti rezultati. Sastavili smo listu najčešćih problema za svaki rečnik. Sastavljena je lista slova i dijakritičkih znakova sa najčešćim problemima u svakom rečniku. Na osnovu ove liste biće generisan dataset sa slikama slova i dijakritičkim znacima kako bi se proširio skup podataka za obuku CNN-a koji se koristi za otkrivanje dijakritičkih znakova.

REZULTATI

Opšte karakteristike obrađenih rečnika

Tabela 1 daje ukupan opis trinaest rečnika obrađenih u aplikaciji SCyDia tako što prikazuje neke od njihovih glavnih karakteristika, relevantnih za OCR, kao što je posedovanje znakova sa dijakritičkim znakovima u naslovnoj reči, znakova sa dijakritičkim obeležjima u citatu, znakova u kurzivu, skraćenica, kao i znakova u superskriptu.

Rečnici	Slova sa dijakriticima u naslovu	Slova sa dijakriticima u citatu	Slova u kurzivu	Sraćenice	Superskript
Bašanović-Čečović 2010	+	+	+	+	+
Boričić Tivranski 2002	+	-	+	+	-
Bukumirić 2012	+	+	+	+	-
Cvetanović 2013	+	+	-	+	-
Cvijetić 2014	+	+	+	+	-
Dalmacija 2004	+	+	+	+	+
Dalmacija 2017	+	+	+	+	+
Đoković 2010	+	-	-	+	-
Rajković Koželjac 2014	+	+	+	+	-
Ristić 2010	+	+	+	+	+
RSGV 2000–	+	+	+	+	-
Stanić 1990–1991	+	+	+	+	+
Zlatković 2014	+	+	+	+	-

Tabela 1: *Opis kompleksnosti rečnika*

Kao što je i bilo očekivano, znakovi sa dijakritičkim znacima u naslovnoj reči prisutni su u svakom od trinaest rečnika. Znakovi sa dijakritičkim znacima u citiranju dokumentovani su u većini rečnika (11 od 13), osim u Boričić Tivranski 2002 i Đoković 2010. Jedanaest od trinaest rečnika ima znakove u kurzivu, osim Cvetanović 2013 i Đoković 2010. Skraćenice (kao što je, recimo, gram), kao i lokacije i izvori, prisutni su u svih trinaest rečnika. Konačno, superskript se nalazi u pet od trinaest rečnika i nedostaje u Boričić Tivranski 2002, Bukumirić 2012, Cvetanović 2013, Cvijetić 2014, Đoković 2010, Rajković Koželjac 2014, RSGV 2000– i Zlatković 2014.

Tačnost OCR obrade

Eksperti su ručno ocenili tačnost OCR obrade. Iako softver SCyDia omogućava poluautomatsko otkrivanje grešaka upoređivanjem malo rotiranih verzija sa originalom, odlučili smo da izvršimo procenu ručno kako bismo bili sigurni da su rezultati evaluacije što tačniji. Poluautomatsko otkrivanje grešaka je veoma korisno za ručno ispravljanje, ali ne možemo biti sigurni da su na ovaj način otkrivene sve greške. Stručnjaci su prebrojali sve greške na stranici, kao i greške u posebnim znakovima (slova sa dijakritikom,

kurzivom i određenim skraćenicama). Želeli smo da vidimo u kojoj meri ovi specijalni znakovi utiču na rezultate OCR-a kako bismo mogli da vidimo koje aspekte bi trebalo da poboljšamo.

Rečnici	**Broj tačno prepo-znatih slova**	**Broj grešaka**	**% tačnosti**	**Broj slova sa dijakri-ticima**	**Greške na dijakri-ticima**	**% tač-nosti**	**% broj grešaka na dijakriti-cima na ukupan broj grešaka**
Cvetanović 2013	1455	2	99,87	107	/	100	/
Đoković 2010	2761	4	99,86	/	/	/	/
Boričić Tivranski 2002	1232	2	99,84	45	/	100	/
Cvijetić 2014	2791	17	99,39	30	/	100	/
Zlatković 2014	3422	33	99,04	263	6	97,8	18,18
Stanić 1990–1991	4394	62	98,59	263	16	94	25,80
Ristić 2010	2938	43	98,54	312	25	92	58,13
Dalmacija 2017	2047	30	98,53	193	15	92,2	50
Rajković Koželjac 2014	3011	47	98,44	175	20	88,6	42,55
Dalmacija 2004	2938	38	98,42	329	5	98,5	13,15
RSGV 2000–	3566	79	97,74	161	14	91,3	17,72
Bašanović-Čečović 2010	2853	61	97,86	355	35	90,1	57,37
Bukumirić 2012	2563	72	97,19	256	6	97,7	8,33

Tabela 2: *Tačnost OCR obrade*

Kao što je prikazano u Tabeli 2, tri rečnika imaju najveći procenat tačnosti (Cvetanović 2013 - 99,87%), (Đoković 2010 – 99,86%), (Boričić Tivranski 2002 – 99,84%). Zajednička karakteristika koju svi dele je nula grešaka kod detekcije dijakritičkih znakova. Pored toga, još jedan rečnik je obrađen bez grešaka u dijakritici (Cvijetić 2014), dakle ukupno četiri rečnika. Kod ostalih rečnika greške na slovima sa dijakritičkim znakovima se većinom vezuju za znakove u kurzivu. Rečnici koji imaju dijakritiku u kurzivu imaju najviše grešaka u dijakritici (Rajković, Koželjac 2014 sa 20 od ukupno 175 znakova sa dijakriticima (88,6% tačnosti), Bašanović Čečović 2010 sa 35 od ukupno 355 (90,1%) i RSGV 2000– sa 14 od ukupno 161 (91,3%)). Specifična vrsta greške u znakovima sa dijakritičkim znakovima prisutna je u većini rečnika – slovo о sa bilo kojom vrstom dijakritike aplikacija SCyDia greškom čita kao ćirilično slovo д. Najveći broj ovih grešaka nalazi se u dva rečnika (Ristić 2010, Dalmacija 2017) gde one čine više od 50% svih grešaka u znakovima sa dijakritikom.

Rečnik	Slova u kurzivu	Grešaka u kurzivu	Broj skraćenica	Broj grešaka u skraćenicama
Cvetanović 2013	/	/	75	/
Đoković 2010	/	/	78	3
Boričić Tivranski 2002	85	1	55	1
Cvijetić 2014	798	17	228	5
Zlatković 2014	755	12	298	6
Stanić 1990–1991	1252	55	267	4
Ristić 2010	828	1	75	/
Dalmacija 2017	669	34	54	2
Rajković Koželjac 2014	231	16	125	/
Dalmacija 2004	820	1	83	/
RSGV 2000–	148	1	452	20
Bašanović-Čečović 2010	627	2	71	2
Bukumirić 2012	483	14	152	17

Tabela 3: *Dalji rezultati dobijeni obradom rečnika u aplikaciji SCyDia*

Rezultati u Tabeli 3 pokazuju da je prisustvo (ili nedostatak) kurziva ključno za ukupan procenat grešaka, posebno ako se kurziv kombinuje sa dijakritičkim znakovima. Rečnici sa najvećim procentom grešaka (Bašanović-Čečović 2010, Dalmacija 201) imaju i znakove u kurzivu i sa dijakritikom. Slično tome, rečnici sa najvećim procentom tačnosti (kao što su Đoković 2010, Cvetanović 2013) nemaju znakove u kurzivu. Ovi rezultati su slični onima koje su Polomac i Lutovac Kaznovac dobili u radu sa OCR-om za srpske srednjovekovne rukopise: „Izuzetno visok procenat grešaka ukazuje da je neophodno obučiti poseban model za automatsko prepoznavanje rukopisa pisanih kurzivnim pismom" (Polomac/Lutovac Kaznovac 2021: 16). Iako je njihov sistem osposobljen za prepoznavanje rukopisa i staroslovenskih slova, zanimljivo je videti da kurziv predstavlja najveći problem, slično našim rezultatima. Takođe je vredno istaći da se značajan procenat grešaka u njihovom istraživanju najčešće odnosi na praznine između reči, nadredna slova i naslove, odnosno dijakritike (Polomac/Lutovac Kaznovac 2021: 23-24).

DALJI PLANOVI

Kada se obradjeni tekst ručno ispravi, koristiće se Python skript da se raščlani i iznese strukturirani rečnik. Trenutno razvijamo OntoLek šemu koja bi bila pogodna za sve rečnike i omogućila nesmetanu integraciju različitih resursa u jednu povezanu strukturu podataka. Na kraju, želimo da napravimo veb aplikaciju pomoću koje bi neki delovi baze podataka bili dostupni široj javnosti, a nekima bi bila potrebna licenca za

pristup, u zavisnosti od autorskih prava rečnika. Takođe, veb aplikacija bi određenom broju korisnika omogućila da uređuje greške koje su možda ostale nakon OCR-a i oskudne ručne ispravke.

ZAKLJUČAK

U ovom radu je predstavljeno SCyDia softversko rešenje izrađeno kao veb aplikacija koja se koristi za OCR obradu tekstualnih stranica pisanih ćiriličnim slovima sa akcentima, dijakritičkim znacima i drugim nestandardnim znakovima (na primer dijalekatski rečnici). SCyDia takođe sadrži modul za poluautomatsku korekciju teksta. Danas, kada se većina rečnika proizvodi u digitalnom obliku, važno je da se ne izgube iz vida oni koji za sada postoje samo u papirnoj formi i koje treba transformisati u digitalni, računarski čitljiv format. Na taj način bi se omogućio novi život nedigitalnim leksikografskim delima, sa krajnjim ciljem stvaranja strukturiranog i indeksiranog materijala koji se može pretraživati i integrisati u različite leksikografske projekte, od naučnih rečnika do popularnijih sadržaja. Ipak, u slučaju srpskog jezika, ovaj krajnji cilj može izgledati nedostižan dok se ne ispune neki osnovni uslovi.

LITERATURA

- Bašanović-Čečović, J. (2010): *Rječnik govora Zete*. Podgorica. [Vocabulary of Zeta (in Cyrillic)]
- Boričić Tivranski, V. (2002): *Rječnik vasojevićkog govora*. Beograd. [Vocabulary of Vasojevići (in Cyrillic)]
- Bukumirić, M. (2012): *Rečnik govora severne Metohije*. Beograd. [Dictionary of the north Metohia (in Cyrillic)]
- Cristea, D./Pădurariu, C./Rebeja, P./Onofrei, M. (2020): *From Scan to Text. Methodology, Solutions and Perspectives of Deciphering Old Cyrillic Romanian Documents into the Latin Script*. In: Knowledge, Language, Models, pp. 38-56; https://profs.info.uaic.ro/~dcristea/papers/Paper%20volume%20Bulgaria-Cristea_etAl.pdf (last access: 15-03-2022)
- Cvetanović, V. (2013): *Rečnik zaplanjskog govora*. Gadžin Han. [Vocabulary of Zaplanje (in Cyrillic)]
- Cvijetić, R. (2014): *Rečnik užičkog govora*. Užice. [Vocabulary of Užice (in Cyrillic)]
- Dalmacija, S. (2004): *Rječnik govora Potkozarja*. Banja Luka. [Vocabulary of Potkozarje (in Cyrillic)]
- Dalmacija, S. (2017): *Rječnik govora Srba zapadne Bosne*. Banja Luka. [Vocabulary of Serbian vernaculars of western Bosnia (in Cyrillic)]
- Đoković, Lj. (2010): *Rječnik nikšićkog kraja*. Podgorica. [Vocabulary of the area of Nikšić (in Cyrillic)]
- Hussain, S. / Niazi, A. / Anjum, U. / Irfan, F. (2014). *Adapting Tesseract for complex scripts: an example for Urdu Nastalique*. In 2014 11th IAPR International Workshop on Document Analysis Systems (pp. 191-195). IEEE.

- Klyshinsky, E. / Karpik, O. / Bondarenko, A. (2020): *A Comparison of Neural Networks Architectures for Diacritics Restoration*. In: Recent Trends in Analysis of Images, Social Networks and Texts. AIST 2020. Communications in Computer and Information Science 1357, pp. 242-253.
- Krstev, C. / Stankovic, R. / Vitas, D. (2018): *Knowledge and rule-based diacritic restoration in Serbian*. In: Computational Linguistics in Bulgaria. Proceedings of the Third International Conference (CLIB), Sofia, 28–29 May 2018. Sofia, pp. 41-51.
- O'Brien, S. / Haddej, D. B. (2012): *Optical character recognition*. Degree of Bachelor of Science. Worcester Polytechnic Institute
- Polomac, V. / Lutovac Kaznovac, T. (2021): *Automatic recognition of Serbian medieval manuscripts by applying the Transcribus software platform: current situation and future perspectives*. In: Zbornik Matice srpske za filologiju i lingvistiku 64/2, pp. 7-26.
- Prepis: *http://www.prepis.org/* (last access: 01-03-2022)
- Rajković Koželjac, Lj. (2014): *Rečnik timočkog govora*. Negotin. [Vocabulary of Timok (in Cyrillic)]
- Raskovnik: *http://raskovnik.org/* (last access: 01-03-2022)
- RSGV (2000–): *Rečnik srpskih govora Vojvodine*. Novi Sad. [Vocabulary of Serbian vernaculars of Vojvodina (in Cyrillic)]
- Rijhwani, S. / Anastasopoulos, A. / Neubig, G. (2020): *OCR post correction for endangered language texts*. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, pp. 5391-5942. https://aclanthology.org/2020.emnlp-main.478/ (last access: 15-03-2022)
- Ristić, D. (2010): *Rječnik govora okoline Mojkovca*. Podgorica. [Vocabulary of the area of Mojkovac (in Cyrillic)]
- Stanić, M. (1990–1991): *Uskočki rečnik*. Beograd. [Vocabulary of Uskoci (in Cyrillic)]
- Stanković, R. / Stijović, R. / Vitas, D. / Krstev, C. / Sabo, O. (2018): *The Dictionary of the Serbian Academy: from the Text to the Lexical Database*. In: Lexicography in global contexts. Proceedings of the 18th EURALEX International congress, Ljubljana, 17–21 July. Ljubljana, pp. 941-949. https://euralex.org/publications/the-dictionary-of-the-serbian-academy-from-the-text-to-the-lexical-database/ (last access: 05-03-2022)
- Stijović, R. / Stanković, R. (2018): *Digitalno izdanje Rečnika SANU: formalni opis mikrostrukture Rečnika SANU*. In: Naučni sastanak slavista u Vukove dane 47/1, pp. 427-440.
- Vitas, D. / Krstev, C. (2015): *Nacrt za informatizovani rečnik srpskog jezika*. In: Naučni sastanak slavista u Vukove dane - Srpski jezik i njegovi resursi: teorija, opis i primene 44/3, pp. 105-116. [Blueprint for the computerized dictionary of the Serbian language (in Cyrillic)]
- Zlatković, D. (2014): *Rečnik pirotskog govora*. Beograd. [Vocabulary of Pirot (in Cyrillic)]

Dr Velibor ILIĆ
MA Lenka BAJČETIĆ
Dr Snežana PETROVIĆ
Dr Ana ŠPANOVIĆ

SCyDia – OCR SOFTWARE FOR SERBIAN CYRILLIC WITH DIACRITICS

Summary

This work presents SCyDia software solution developed as a web application that is used for OCR processing of text pages written in Cyrillic script with accents, diacritic and other non-standard marks (dialect dictionaries for example). SCyDia also contains a module for semi-automatic text correction. Today, when most of the dictionaries are made in digital format, it is important not to forget those made only in paper form which should be transformed into digital, machine-readable format. In that way, non-digital lexicography texts would have a new life, with the end goal of creating structured and indexed material that can be searched and integrated into different lexicography projects, from scientific dictionaries to popular contents. However, in the case of the Serbian language, this end goal may seem unattainable until some basic conditions are met.